Classification of Student Majors with C4.5 and Naive Bayes Algorithms (Case Study: SMAN 2 Bekasi City)

Antonius Yadi Kuntoro, Hermanto, Lasman Effendi, Ridatu Oca Nitra, Hermanto Wahono, Riza Fahlapi, Ferry Syukmana, Sumpena, Yuma Akbar, Taufik Asra, Sfenrianto

ABSTRACT

School majors conducted in high school are based on interests and these have a goal to provide opportunities for learners to develop the competence of attitudes, skills competence of learners in accordance with interests, talents, and academic ability in a group of scientific subjects.In this research, the researcher uses two algorithm models that is a comparison between the C4.5 algorithm and also the Naive Bayes algorithm. In this study, the data used is the results of school entrance test data and also the data from psychological results for students who have been declared passed the entrance test school SMAN 2 Bekasi City academic year 2018/2019. By comparison of two data mining classification algorithm, can be proved with accuracy result and AUC value from each algorithm that is for Naive Bayes accuracy = 76,43% and AUC value = 0,846, while for algorithm C4.5 accuracy = 70,29% and AUC value = 0.738.

**Keywords**: Algorithm C4.5; Naïve Bayes; Student majors

Introduction

In the new curriculum, the majors are conducted at the beginning of the school entry, namely in class X. Changes in the curriculum are intended to enable the adjustment of educational programs to educational units with the potential conditions and peculiarities in the student area (Kemdikbud regulations). The possibility that will occur if students experience errors in majors is the low student achievement or can cause incompatibility with the direction chosen by the previous student or student (Education, Culture, & Indonesia, 2013).

In the process of education in school, differences in each student must be considered because it can determine the good and bad of student learning achievement. The basic purpose of the school is to develop all the talents and abilities of students during the education process. Individual differences between students in schools include differences in cognitive abilities, achievement motivation, interest, and creativity. And with these individual differences, the function of education is not only in the teaching and learning process, but also counseling, so that the selection and placement of majors for students must be based on the individual capacity as a student. (Bahar, 2011).

Placement of students according to their capacity or often referred to as student majors in secondary school is determined by academic abilities supported by factors of interest. It is because of the characteristics of science require the same characteristics from those who learn it. Thus, students who study science that is in accordance with their personality characteristics will feel happy when studying science. Interest can affect the quality of student learning outcomes in a particular field of study. A student who is interested in Mathematics, for example, will focus more on the Mathematics field than others. Because of the concentration of intensive attention to the material, students will study harder and achieve desired achievements (Bahar, 2011).

The incompatibility of students' competencies towards the majors they take is that the decision maker must really consider the criteria that have been set in the decision-making department. This will affect the success of student learning. Determination of majors is a problem experienced by students who want to continue their education to a higher level. (Hertyana, 2018)

Based on the results of interviews with the school in determining the majors of student specialization conducted by SMAN 2 Bekasi City, there are only two majors, namely majors in Natural Sciences and Social Sciences. Determination of these majors is considered based on the entrance test scores of students of class X who pass the entrance examination, as well as student’s interests and talents seen from the psychological test results after they have passed the entrance examination. However, there are still lots of students who ignore their abilities or desires because they prefer following friends’ choice since they want to be in the same class or other possible factors.

However, Classification and Clustering activities carried out by humans still have some limitations, especially in the ability of humans to accommodate the amount of data they need to process. In addition, errors can also occur due to inaccuracies carried out. One way to overcome this problem is to use data mining techniques to process data into strategic information sources using Classification and Clustering methods. Data mining can help an organization that has abundant data to provide information that can support decision making. (Nugroho, 2017)

Study of Literature

1. *Data Mining Process*

Data Mining is a process that must be in accordance with the procedures of the process itself, this process is called The Cross-Industry Standard Process for Data Mining (CRISP-DM). The CRISP-DM methodology is an effort to standardize the data mining process. The process model in CRISP-DM provides an overview of the life cycle of a data mining project. The process contains the project phase, each task and the relationship between tasks.

In CRISP-DM, there are six phases that are interconnected to describe the data mining process, namely: business understanding, understanding data, data preparation, modeling, evaluation, and deployment (Maimon&Rokach, 2010 p.1032).

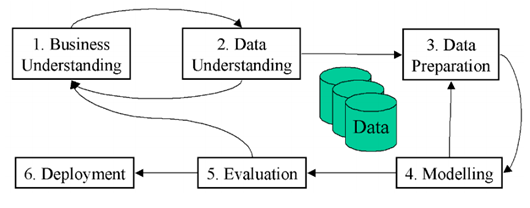


Figure 2.1 Process CRISP-DM

Source: (Maimon & Rokach, 2010 p.1032)

1. *Algorithm C4.5*

The C4.5 algorithm was introduced as an improved version of ID3. In ID3, induction of decision trees can only be done on features of the categorical type (nominal and ordinal), while the numeric type (interval or ratio) cannot be used. The improvement that distinguishes C4.5 algorithm from ID3 is that it can handle features with numeric types, pruning decision trees, and deriving rule sets. The C4.5 algorithm also uses gain criteria in determining the features that become node breakers in the induced tree (Prasetyo, 2014).

What is important in the induction of decision trees is how to declare the testing conditions on the node. There are 3 important groups in node testing requirements:

* 1. Binary Features

Features that only have two different values are called binary features. Testing conditions when this feature becomes a node (root or internal) only have two options. Examples of solutions are presented in Figure 2.2

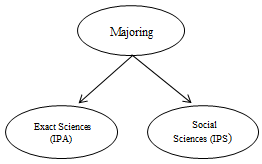


Figure 2.2 Terms of testing binary features (Prasetyo, 2014)

* 1. Categorical type features

For features that have categorical type values (nominal and ordinal) can have several different values. An example is the 'weather' feature has 3 different values, and this could have many combinations of test testing requirements. In general, there are two, namely binary splitting and multi splitting.

* 1. Numerical type features

For numerical features, the testing requirements in nodes (root or internal) are expressed by comparison testing (A <ѵ) or (A ≥ ѵ) with binary results, or for multi with results in the form of a range of values in the form ѵᵢ ≤ A <ѵᵢ + 1 , for i = 1, 2, ..., k. In the case of binary solving, the algorithm will check all possible solving positions ѵ and choose the best position ѵ. For multi-methods, the algorithm must check all possible continuous values.

1. *Naïve Bayes*

Naive Bayes is a simple probabilistic-based prediction technique that is based on the application of the Bayes theorem (Bayes rule) with a strong (naive) independence assumption. In other words, the Naive Bayes model used is an "independent feature model". Naive Bayes is one of the most effective and efficient inductive learning algorithms for machine learning and data mining. Naive Bayes' performance is competitive in the classification process even though it uses attribute independence assumptions (there is no connection between attributes). The independent assumption of this attribute on actual data is rare, but even though the independent assumptions of these attributes are violated, the performance of classifying NaiveBayes is quite high. This is evidenced in various empirical studies (Musthofa, 2016).

1. *Rapid Miner*

RapidMiner is a software that is open (open source). RapidMiner is a solution for analyzing data mining, text mining and prediction analysis. RapidMiner uses a variety of descriptive and predictive techniques to provide insight to users so that they can make the best decisions. RapidMiner has approximately 500 data mining operators, including operators for input, output, data preprocessing and visualization. RapidMiner is stand-alone software for data analysis and as a data mining machine that can be integrated into its own products. RapidMiner is written using Java language so that it can work on all operating systems (Aprilia et al: 2013).

1. *Confusion Matrix*

Confusion Matrix (Gorunescu, 2010) is a visualization tool commonly used supervised learning. Each column in the matrix is an example in a prediction class, while each row represents an actual class event. One advantage of The Confusion Matrix is that it is easy to know if data exists between two classes (mislabeling). The Confusion Matrix contains information about actual and predicted conditions in the classification system. System performance like this is usually evaluated using data on the matrix.

Evaluation of the classification model is based on testing to estimate the right and wrong objects, the sequence of tests tabulated in the confusion matrix where the predicted class is displayed at the top of the matrix and the class observed on the left side. Each cell contains a number that shows how many actual cases of the class are observed to be predicted (Khafiizh, 2012).

1. *ROC Curve*

ROC Curve (Gorunescu, 2010) is another way to test the performance of classifiers. A ROC chart is a plot with a false positive level (FP) on the X-axis and a true positive level (TP) on the Y-axis. Point (0.1) is a perfect classification that classifies all positive and negative cases correctly, because the positive rate is wrong (FP) is 0 (none), and the true positive level (TP) is 1. Point (0,0) is a classification that predicts each case to be negative, while point (1,1) corresponds to a classification that predicts each case becomes positive. Point (1.0) is an incorrect classification for all classifications.

1. *K-Fold Cross Validation*

K-Fold Cross Validation is a validation technique (Witten, Frank, and Hall 2007: p152) which divides data into k sections and then each classification process will be carried out. Using K-Fold Cross Validation will be carried out as many experiments as K. Each experiment will use one testing data and k-1 part will become training data, then the testing data will be exchanged with one training data so testing data will be obtained for each experiment different. Training data is data that will be used in conducting learning while testing data is data that has never been used as a learning and will function as test data for the truth or accuracy of learning outcomes. In this study, the value of k used amounted to 10 or 10-Fold Cross-Validation.

8. *T-Test*

The T-Test is a method of testing hypotheses using one individual (research object) using two different treatments. Despite using the same object, the sample is still divided into two, namely data with the first treatment and data with the second treatment. Performance can be known by comparing the conditions of the first research object and the condition of the object in the second study (Khafiizh, 2012).

1. *Classification*

Classification can be defined in detail as a job that conducts training or learning on the target function ƒ which maps each vector (feature set) χ into one of the numbers of available classy labels. The training workshop will produce a model which is then stored as memory. The model in classification has the same meaning as a black box, where there is a model that receives input and is then able to think about the input and provide answers as outputs from the results of his thoughts (Prasetyo, 2014).

Methodology

For Data Mining research, there has been a standard methodology called CRISP-DM or Cross-Industry Standard Process for Data Mining. CRISP-DM is a collaboration of several companies, including Daimler-Benz, OHRA, NCR Corp, and SPSS Inc. which started since 1999 (North et al., 2014).

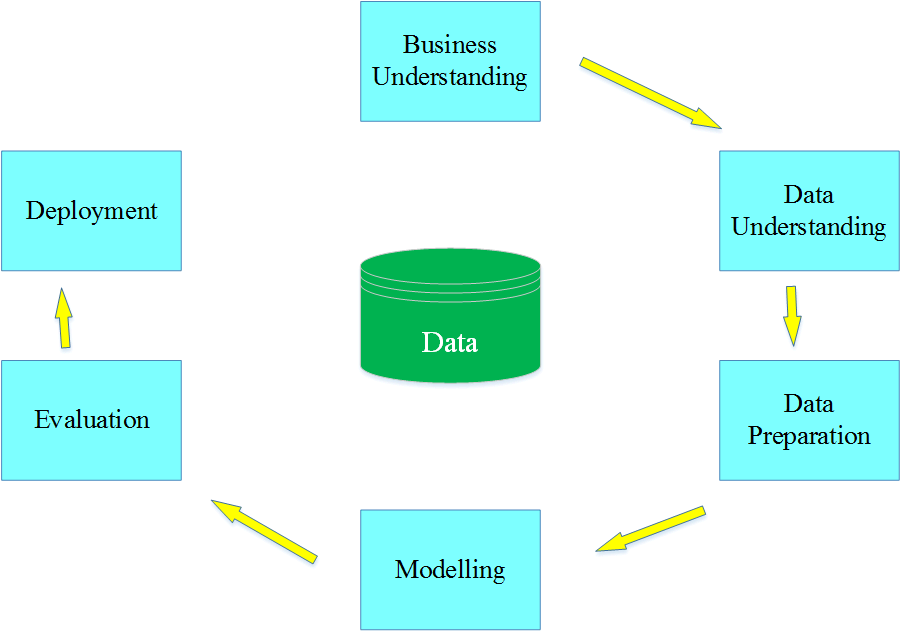


Figure 3.1 CRISP-DM (North, 2012)

CRISP-DM has six stages (North, 2012), namely:

1. Business Understanding

In this first stage, knowledge needs to be defined in the form of general questions, such as how to increase profits, how to anticipate product defects, and so on.

1. Data Understanding

This second stage aims to collect, to identify, and to understand the data assets that we have. The data must be verified by truth and reliability.

1. Data Preparation

This stage includes some activities, such as cleaning data, reformatting data, reducing the amount of data, etc. These aim to prepare data to be consistent in the format needed.

1. Modeling

The model is a computational representation of the results of observations that are the result of searching and identifying the patterns contained in the data.

1. Evaluation

The Evaluation aims to determine the usefulness of the model that we have made in the previous step.

1. Deployment

Deployment is a time when the results of all previous stages are used in real terms.

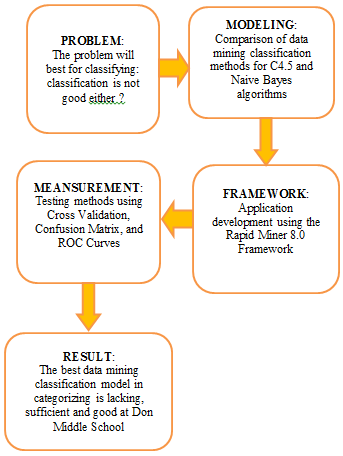
1. *CRISP-DM Stages*

The research method used in this study is to use the experimental method. This study aims to compare and evaluate data mining classification algorithms in determining majors in at SMAN 2 Bekasi city. In designing this experimental research method, researchers used a standard research method used in data mining, namely Cross Industry Standard Process for Data Mining (CRISP-DM) which consists of 6 phases with the steps are Business Understanding, Understanding Data, Data Preparation, Modeling, Evaluation and Deployment (Larose, 2005).

1. *Framework*

In completing this study, the author makes a framework of thinking that is useful as a guideline or reference in this study so that this research can be done consistently. This study consists of several stages as seen in Figure 3.2 frame of mind. The problem in this study is whether the C4.5 and Naive Bayes algorithms can be applied to the determination of the science class majors and the IPS class at SMAN 2 Bekasi city and which algorithms will provide the best model for classifying the majors in SMAN 2 Bekasi city.

For this reason, a model is made using the C4.5 and Naive Bayes algorithms to solve the problem, then test the two performance models. After testing the two models formed, 10 fold cross validation will be tested. The accuracy of the two models that have been formed will be measured by using a confusion matrix, while the Under Curve Area (AUC) will be measured using ROC Curve. To develop the application (deployment) based on the model that has been made, the Rapid Miner 8.0 tools are used. The following is a description of the framework that has been carried out as follows:



**Figure 3.2** Framework

Result and Performance Analysis

1. *Implementation Of The Methodology*

Based on the research methodology described in chapter III, the following methodology implementation was carried out in this study.

1. *Research Methods*

In this study, the authors conducted a study using the CRISP-DM research method (Cross-Standard Industry for Data Mining). The Stage of Crisp-Dm consists of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment (Brown, 2014).

1. *Business Understanding*

The first stage of CRISP-DM is Business Understanding which is to understand student data at SMAN 2 Bekasi City in determining its special majors new class X students entering the determinants of the majors only refer to the results of the psychological test only, hence researchers here propose a new attribute that is the result of test entrance tests for X grade students used with the results of psychological tests to determine the results of student majors. Which will be tested by comparing two algorithms namely C4.5 and Naive Bayes.

1. *Data Understanding*

The second stage of CRISP-DM is Data Understanding. At this stage, the author examines the data of new prospective students of SMAN 2 Bekasi City in knowing the 2018/2019 lesson. The data taken is the results of the school entrance test data, and psycho-test of new students of SMAN 2 Bekasi City. The contents of entrance test result are subject values that were tested include Bahasa Indonesia, English, Mathematics, Physics, Biology, and Religion. The data used in this study is the data from the X class entrance test results at SMAN 2 Bekasi city and also the psychological test results of prospective class X students in the 2018-2019 academic year with a total of 148 students from two science classes and three social studies classes. Where there were as many as 60 female students and 88 female students. There were 83 students in the science class and 65 in the social studies class. The following is the initial data before the value of the numbers is converted, see Table 4.1 below:

Table 4.1 Examples of Test Exam Values

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No | Bahasa | English | Mathematics | Physics | Biology | Religion |
| 1. | 17 | 9 | 4 | 13 | 13 | 18 |
| 2. | 13 | 7 | 9 | 9 | 12 | 14 |
| 3. | 17 | 9 | 15 | 16 | 16 | 23 |
| 4. | 17 | 5 | 14 | 17 | 17 | 21 |
| 5. | 19 | 16 | 7 | 16 | 11 | 19 |
| 6. | 17 | 13 | 5 | 16 | 16 | 21 |
| 7. | 17 | 14 | 15 | 17 | 14 | 23 |
| 8. | 15 | 8 | 11 | 10 | 14 | 21 |
| 9. | 18 | 16 | 13 | 11 | 13 | 17 |
| 10 | 17 | 10 | 6 | 17 | 17 | 20 |

Data for class X students in the academic year 2018/2019

**Table 4.2** Table Number of Students

|  |  |
| --- | --- |
| **Class** | **The number of students** |
| X IPA 1 | 33 students |
| X IPA 2 | 31 students |
| X IPS 1 | 30 students |
| X IPS 2 | 28 students |
| X IPS 3 | 26 students |

Table 4.2 above is the amount of data for class X students in the 2018/2019 academic year received at SMAN 2 Bekasi City in the IPA and IPS classes which the data will be processed.

Class X psychological test results for school year 2018/2019

Table 4.3 Sample Table of psychological test results

|  |  |  |
| --- | --- | --- |
| **No** | **Gender** | **Majoring** |
| **1** | F | IPA |
| **2** | M | IPA |
| **3** | M | IPA |
| **4** | F | IPA |
| **5** | F | IPA |
| **6** | F | IPS |
| **7** | F | IPS |
| **8** | F | IPS |
| **9** | M | IPS |
| **10** | M | IPS |

1. *Data Preparation*

The third stage of CRISP-DM is Data Preparation. This stage prepares data so that it can be processed at the next stage. This stage adjusts the attributes of the data table that will be processed using the Rapid Miner 8.0 Framework. Several tables that have been described will be processed in this study and in this phase, the author creates a new table combining the tables into one table. The data in Table 4.5 below is the result of merging from the previous data tables that will be included in the process using the Rapid Miner 8.0 Framework.

Table 4.4 Student data table

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | Bahasa | English | Mathematics | Physics | Biology | Religion | Gender | Majoring |
| 1 | 17 | 9 | 15 | 15 | 16 | 23 | F | IPA |
| 2 | 17 | 5 | 14 | 17 | 17 | 21 | M | IPA |
| 3 | 19 | 16 | 7 | 16 | 11 | 19 | F | IPA |
| 4 | 17 | 13 | 5 | 16 | 16 | 21 | F | IPA |
| 5 | 17 | 14 | 15 | 17 | 14 | 23 | F | IPA |
| 6 | 18 | 11 | 10 | 12 | 14 | 21 | M | IPS |
| 7 | 16 | 10 | 9 | 12 | 9 | 20 | F | IPS |
| 8 | 18 | 9 | 9 | 12 | 12 | 14 | F | IPS |
| 9 | 13 | 7 | 7 | 13 | 13 | 20 | M | IPS |
| 10 | 20 | 13 | 9 | 15 | 6 | 16 | M | IPS |

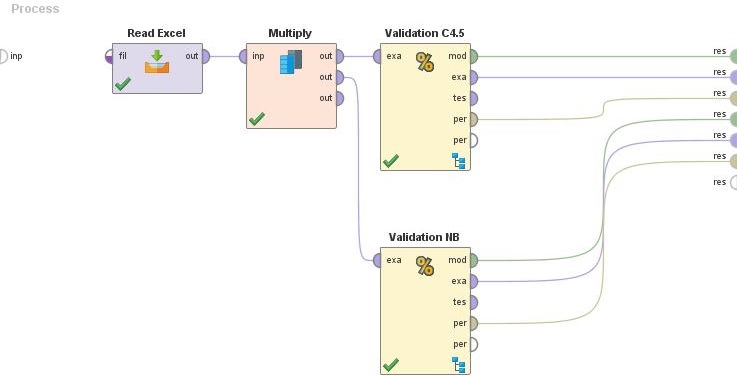
In the student data table above, the researcher reprocessed the data by converting redundant values or values that were too diverse into smaller groups to facilitate the formation of the model. For example, a value with a number of ≤ 2 is categorized sufficient and ≥ 17 is categorized as very good, Table 4.5 is a table of categorization of attributes can be seen below:

Table 4.5 Attribute category

|  |  |
| --- | --- |
| Score | Category |
| 2 – 9 | Enough |
| 10 – 16 | Good |
| 17 – 24 | Very Good |

1. *Modeling*

The fourth stage in CRISP-DM is Modeling. At this stage, the dataset that was made in the previous stage is used as input for the classification algorithm, which is used as a training dataset. In this study, two types of classification algorithms will be used, namely C4.5 and Naive Bayes. The following is the design process that is used along with the description:



1. *Evaluation*

The fifth stage of the CRISP-DM method is Evaluation aiming to determine the usefulness of the model we have succeeded in making in the previous step. For evaluation, 10-fold cross-validation is used. The following is the design process used.

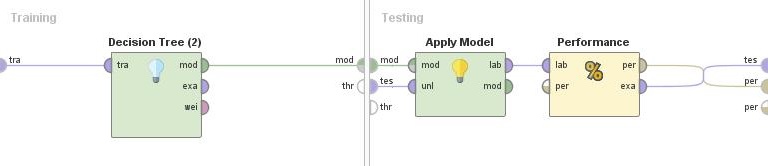
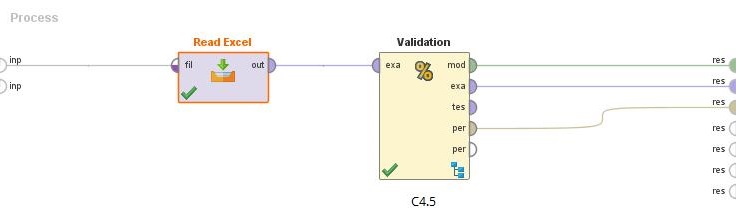


Figure 4.2 Testing the C4.5 Algorithm Model

After testing with the model above, the results formed will look like in Figure 4.4 below:

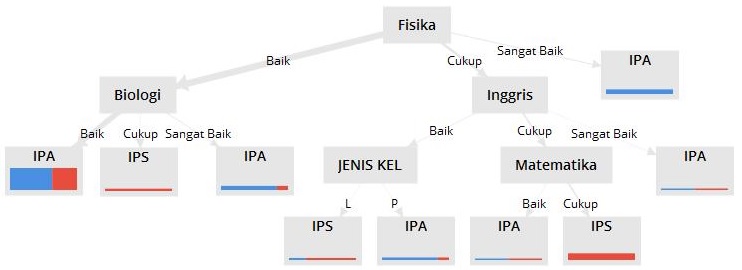


Figure 4.3 C4.5 Algorithm Decision Tree Model

The following are the rules of the C4.5 algorithm tree decision model, as follows:

1. R1 : Physics = Good and Biology = Good then Science

2. R2 : Physics = Good and Biology = Enough Social Studies

3. R3 : Physics = Good and Biology = Very Good then Science

4. R4 : Physics = Very Good then Science

5. R5 : Physics = Enough and English = Good and Type Kel = L then IPS

6. R6 : Physics = Enough and English = Good and Type Kel = P then Science

7. R7 : Physics = Enough and English = Enough and Mathematics = Good then Science

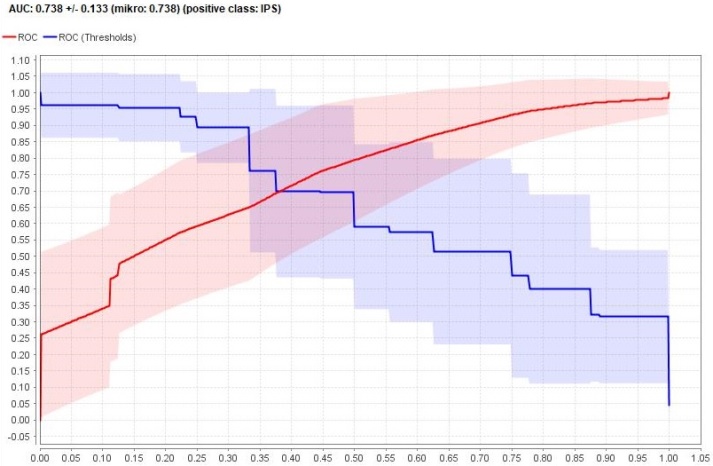
8. R8 : Physics = Enough and English = Enough and Mathematics = Enough then IPS

9. R9 : Physics = Enough and English = Very Good then Science

Based on the results of testing using Rapid Miner, a decision tree and model rules are obtained as seen above with attributes such as Name, Bahasa, English, Mathematics, Physics, Biology, Religion, IQ, Gender, and Department. However, in the decision tree above, not all attributes appear because the attribute has a small gain value.

The results of the model testing that has been done aims to get the results of accuracy and Under Curve Area (AUC) and to get the results of the ROC graph with the value of Area Under Curve (AUC) of 0.738 with accuracy performance, namely Fair Classification, as seen in Figure 4.5 below:

a. ROC Curve (AUC)



**Figure 4.4** AUC value in C4.5 Algorithm

From the results of the tests that have been done with the C4.5 Algorithm model, the accuracy value is 70.29%, as shown in Table 4.6 below:

b. Accuray

**Table 4.6** Model for C4.5 Algorithm

Accuracy:70.29% +/- 9.72% (mikro: 70.27%)

|  |  |  |  |
| --- | --- | --- | --- |
|  | True IPA | True IPS | Class Precision |
| Pred. iPA | 74 | 35 | 67.89% |
| Pred. IPS | 9 | 30 | 76.92% |
| Class Recall | 89.16% | 46.15% |  |

The number of True Positives (TP) is 74 classified records as selected IPA and False Negative (FN) as many as 9 records classified as selected IPS. 35 records for Positive False are classified as selected IPA, and 30 records for True Negative are classified as selected IPS. Based on Table 4.1 above, it shows that the level of accuracy using the C4.5 algorithm is 70.29%. Similar to the evaluation process in the C4.5 algorithm above, the Naive Bayes algorithm is also evaluated as below:

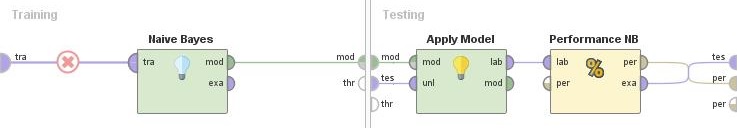
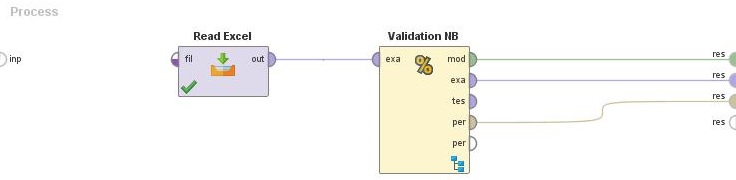


Figure 4.5 Testing the Naive Bayes Algorithm Model

The results of the model testing that has been done aims to get the results of accuracy and Under Curve Area (AUC) and to get the results of the ROC chart with the value of Area Under Curve (AUC) of 0.846 with accuracy performance, namely Good Classification, as seen in Figure 4.6 below:

a. ROC Curve (AUC)

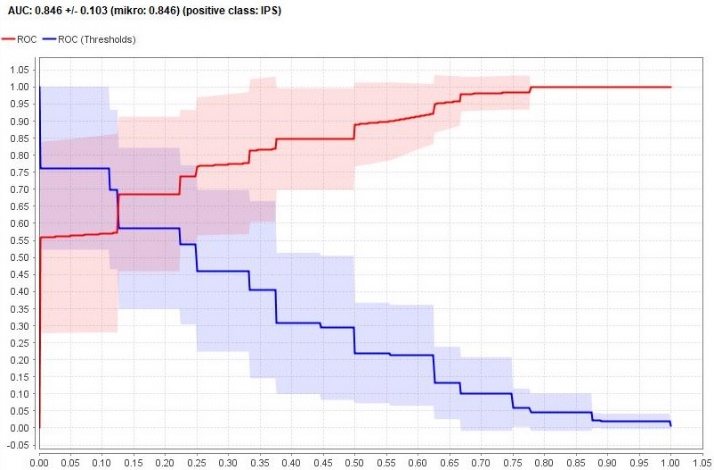


Figure 4.6 The AUC value in the Naive Bayes Algorithm

From the results of testing that has been done with the Naive Bayes Algorithm model, the accuracy value is 76.43%, as shown in Table 4.8 below:

b. Accuracy

Table 4.7 Model for Naive Bayes Algorithm

Accuracy:76.43% +/- 12.38% (mikro: 76.35%)

|  |  |  |  |
| --- | --- | --- | --- |
|  | True IPA | True IPS | Class precision |
| Pred. IPA | 69 | 21 | 76.67% |
| Pred. IPS | 14 | 44 | 75.86% |
| Class recall | 83.13% | 67.69% |  |

The number of True Positives (TP) is 69 records classified as selected IPA and False Negative (FN) as many as 21 records classified as selected IPS. In addition, 21 records for Positive False are classified as selected IPA, and 44 records for True Negative are classified as selected IPS. Based on Table 4.6 above, it shows that the level of accuracy using the Naive Bayes algorithm is 76.43%.

1. *Deployment*

This stage is the last stage in the standard modeling in data mining (CRISP-DM). In this stage, the report will be made in the form of writing the results of thesis and journal research from the introduction to conclusions, as well as creating a Graphical User Interface (GUI) so that later, users who use the results of this study can interact and apply it easily.

1. *Comparative Performance*

Based on the results of the analysis of each of the algorithm tests above, the results can be summarized as in Table 4.8 below:

Table 4.8 Comparison of Performance Algorithms

|  |  |  |
| --- | --- | --- |
|  | C4.5 | Naive Bayes |
| ACCURACY | 70,29% | 76,43% |
| AUC | 0,738 | 0,846 |

From the results of the performance comparisons of the two algorithms above, the test results for Naive Bayes have higher accuracy than the C4.5 algorithm. The accuracy value for the Naive Bayes algorithm model is 76.43% and the accuracy value for the C4.5 algorithm model is 70.29% with an accuracy difference of 6.14%.

1. *Analysis of Comparative Results*

Looking at the results of calculations in Table 4.7 above and by applying the accuracy of the Area Under Curve (AUC) performance classification, the results of this study can be divided into two classifications namely Fair Classification for C4.5 algorithm with AUC of (0.738) and Good Classification for Naive Bayes algorithm with an AUC value of (0.846).

After testing with 10 Fold Cross Validation in this study, it was tested again using T-Test to test the truth and falseness of the two models. In this test, a comparison between two algorithms will be carried out, namely C4.5 and Naive Bayes. To get the results of the T-Test statistics calculation on the C4.5 and Naive Bayes algorithms, it will be seen in Table 4.9 below:

Table 4.9 Test results for C4.5 and Naive Bayes T-Tests

|  |  |  |
| --- | --- | --- |
| A | B | C |
|  | 0.729 +/- 0.089 | 0.764 +/- 0.124 |
| 0.729 +/- 0.089 |  | 0.469 |
| 0.764 +/- 0.124 |  |  |

Based on Table 4.10 above, it can be analyzed that the C4.5 and Naive Bayes algorithms have insignificant differences in values, and have a probability of> 0.05, which is 0.469.

Conclusion

Based on the results of the research, it can be concluded that: from the results of tests that have been produced, the accuracy value for the determinants of student majors at SMAN 2 Bekasi city with a comparison of two data mining classification algorithms, can be proved by the results of accuracy and AUC values of each The algorithm, for Naive Bayes accuracy = 76.43% and AUC = 0.846, while for C4.5 algorithm accuracy = 70.29% and AUC = 0.738.Suggestions given by the author to perfect the results of this study are, with the application of the Naive Bayes algorithm, it is expected to be able to provide solutions for students of SMAN 2 Bekasi City in determining the majors according to the abilities, interests, and talents of students. The following suggestion is to add several optimization algorithms such as Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) as well as other algorithms to increase the level of accuracy especially in determining high school majors so that they can be more useful for students in the majors.

References

Aprilia, Dennis, Baskoro, Donny Aji, Ambarwati, Lia, Wicaksana, I WayanSimri. (2013). Belajar Data Mining dengan Rapid Miner. Jakarta

Bahar, (2011). Penentuan Jurusan Sekolah Atas Dengan Algoritma Fuzzy C-Means. Universitas Dian Nuswantoro Semarang.

Gorunescu, Florin (2011). Data Mining: Concept, Model, And Techniques. Verlag Scoring Berlin Heidelberg: Springer.

Hastuti, Khafiizh. (2012). Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif. Intertational Standart Book Number, 979, 26-0255.

Musthofa, Mufid. (2015). Pengembangan system pendukung keputusan penjurusan bagi siswa baru menggunakan metode naive bayes. Polteknik Negeri Malang..

North, Matthew (2012). Data Mining for The Masses. A Global Text Project Book.

Prasetyo, Eko. (2014). Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab. Yogyakarta: Andi Offset.

Pendidikan, M., Kebudayaan, D. A. N., & Indonesia, R. (2013). Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia. No 81A Tahun 2013.

Pendidikan, M., Kebudayaan, D. A. N., & Indonesia, R. (2014). Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia. No 64 Tahun 2014.

Witten, Ian H., (2011). Data Mining Practical Machine Learning Tools and Techniques.

Hertyana, H. (2018). ANALISA PENENTUAN JURUSAN PADA SMA. KARTIKA VIII-1 MENGGUNAKAN METODE FUZZY INFERENCE SYSTEM MAMDANI. *Jurnal Ilmu Pengetahuan Dan Teknologi Komputer*.

Nugroho, Y. S. (2017). Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*. <https://doi.org/10.23917/khif.v1i1.1175>

Oded Maimon, LiorRokach, (2010), 2nd Edition Data Mining And Knowledge Doscovery Handbook

*Proceeding of Geosynthetics Conference*, Nashville, Tennesee, Vol. 2, (pp.433-446).

Hardiyatmo, H.C. (2006). *Mekanika Tanah 1*.Yogyakarta: Gadjah Mada University Press.

Kaniraj, S.R. & Havanagi, V.G. (2001). Behavior of cement-stabilized fiber-reinforced fly ash-soil mixtures. *Journal of Geotechnical and Geoenviromental Engineering*, *127*(7), 574-584.

Penulis:

Antonius Yadi Kuntoro

STMIK Nusa Mandiri, Jakarta,

Email: antonius.aio@nusamandiri.ac

Hermanto

STMIK Nusa Mandiri, Jakarta,

Email: [Hermanto.hmt@bsi.ac.id](mailto:Hermanto.hmt@bsi.ac.id)

Lasman Effendi  
STMIK Nusa Mandiri, Jakarta,

Email: [lasman.lef@bsi.ac.id](mailto:lasman.lef@bsi.ac.id)

RidatuOca Nitra

STMIK Nusa Mandiri, Jakarta,

Email: Ridat.rdo@bsi.ac.id

HermantoWahono

STMIK Nusa Mandiri, Jakarta,

Email: [mr.h3rm4n@gmail.com](mailto:mr.h3rm4n@gmail.com)

Riza Fahlapi  
STMIK Nusa Mandiri, Jakarta,

Email: [riza.fahlapi@gmail.com](mailto:riza.fahlapi@gmail.comd)

Ferry Syukmana

STMIK Nusa Mandiri, Jakarta,

mail: 14002191@nusamadiri.ac.id

Sumpena

STMIK Nusa Mandiri, Jakarta,

Email: sumpenasuhandi@gmail.com

Yukma Akbar

STMIK Nusa Mandiri, Jakarta,

Email: yuma.pjj@gmail.com

TaufikAsra

STMIK Nusa Mandiri, Jakarta,

Email: Taufik.tas@bsi.ac.id

Sfenrianto

STMIK Nusa Mandiri, Jakarta,

Email: sfenrianto@binus.edu